Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments

Tsyplakov, Alexander

Department of Economics, Novosibirsk State University

18 March 2013

Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments

Alexander Tsyplakov

Department of Economics, Novosibirsk State University

March 18, 2013

Abstract

The paper provides an overview of probabilistic forecasting and discusses a theoretical framework for evaluation of probabilistic forecasts which is based on proper scoring rules and moments. An artificial example of predicting second-order autoregression and an example of predicting the RTSI stock index are used as illustrations.

Key words: probabilistic forecast; forecast calibration; probability integral transform; scoring rule; moment condition.

JEL classification: C53; C52.

1 Introduction

In the recent years increasing emphasis in the forecasting literature is made on various probability forecasts. The initial impetus for the development of probabilistic forecasting was made by meteorological research. Conventional deterministic or categorical forecasts (in the form of "tomorrow it will rain") have many potential pitfalls (e.g. Murphy and Winkler, 1984). It is desirable that the forecaster reported not only some plausible level of predicted variable (i.e. point forecast), but also the associated uncertainty and probabilities of different scenarios. "With the availability of uncertainty information, users—each with their own sensitivity to costs and losses and with varying thresholds for taking protective action—could better decide for themselves whether to take action and the appropriate level of response to hydrometeorological situations" (National Research Council (U.S.), 2006, p. 13). This pertains both to the simplest daily activities (e.g., how to dress today) and to natural disasters with devastating effects. A detailed review of the development of probabilistic forecasting in meteorology can be found in Murphy and Winkler (1984).

Clearly, these considerations fully apply to forecasting of economic variables: it is important for users to have information on the degree of forecast uncertainty and probabilities of different scenarios. That is why the probabilistic forecasts are becoming increasingly popular among economists.

Perhaps the most well-known density forecast is the Bank of England inflation forecast (Britton et al., 1998; Clements, 2004). The forecasts are published by the Bank of England since 1996. The forecasts are quarterly, with horizons from one quarter to two years. The density forecast is given by a two-piece normal distribution combined from two scaled halves of the usual normal distribution; this kind of predictive distribution captures asymmetry. Graphical representations of the Bank of England forecasts are published in the form of so-called *fan charts* composed of interval forecasts. A detailed description of corresponding procedures can be found in Britton et al. (1998).

Another well-known series of probabilistic forecast is provided by the Survey of Professional Forecasters conducted by the Federal Reserve Bank of Philadelphia (Engelberg et al., 2009; Diebold, Tay and Wallis, 1999). Among other things, the survey experts are asked to assess probabilities that the predicted macroeconomic indicator falls in specified ranges. This procedure does not allow the experts to specify the complete distribution function which leads to forecast discretization, i.e. the forecasts are histogram-like.

The practice of providing point forecasts without specifying probabilities and probability distribution functions still dominates, but it is gradually realized that interpretation and use of point forecasts may be associated with serious practical difficulties. Particularly many problems arise when the forecaster is not put to a specific task, i.e. there is no indication of what particular feature of the forecast distribution (the mean, mode, median) is to be predicted or what scoring function is to be used to measure the success of the forecast (for example, absolute prediction error or squared prediction error). Some of these issues are highlighted in Engelberg et al. (2009) and Gneiting (2011).

One can describe the simplest scheme of decision-making based on probabilistic forecasts as follows. The forecast user chooses some action a. The consequences depend on a realization y of a random variable Y. If the preferences of the forecast user are described by a utility function[1] $u(a, y)$ and F is a probabilistic forecast of Y in the form of a probability distribution function, then the best action $a(F)$ is given by (e.g. Pesaran and Skouras, 2002)

$$a(F) \in \arg\max_a \mathsf{E}[u(Y, a)] \qquad \text{for } Y \sim F.$$

(By abuse of notation F is either a non-random or random distribution function depending on the context).

One can say that F_1 is better than F_2 if it leads to a greater expected utility, that is,

$$\mathsf{E}u(Y, a(F_1)) > \mathsf{E}u(Y, a(F_2)).$$

This provides economic foundation for the theory of evaluation of probabilistic forecasts. Note that in this formulation F_1 and F_2 are function-valued random elements.

On the technical side, the key to many theoretical results mentioned in this paper is the substitution property of the conditional expectation. Consider a measurable real-valued function $b(x_1, x_2)$ and two random elements X_1 and X_2 such that $\mathsf{E}|b(X_1, X_2)| < \infty$. If X_2 is measurable with respect to a sub-sigma-algebra $\mathcal{A}$, then X_2 can be treated as fixed inside the conditional expectation with respect to $\mathcal{A}$. That is, $\mathsf{E}[b(X_1, X_2)|\mathcal{A}] = B(X_2)$, where $B(x_2) = \mathsf{E}[b(X_1, x_2)|\mathcal{A}]$. Unfortunately, the general result is not readily available from the literature on probability theory. A well-known particular case is $b(x_1, x_2) = x_1 x_2$ when we have $\mathsf{E}[X_1 X_2|\mathcal{A}] = \mathsf{E}[X_1|\mathcal{A}]X_2$. Another variant relates to independent X_1 and $\mathcal{A}$ (Bhattacharya and Waymire, 2007, Theorem 2.7(ℓ)).

This paper discusses a theoretical framework for comparison and assessment of probabilistic forecasts. Section 2 introduces the notions of proper scoring rule and various modes of calibration and discusses their properties and relationships among them. Section 3 discusses testable conditions for the different modes of calibration. Sections 4 and 5 provide illustrative examples. Section 6 concludes.

[1] In forecasting theory one often uses expected loss minimization instead of expected utility maximization.

2 Key concepts, definitions and properties

2.1 Scoring rules

A *scoring rule* is a function $S(F,\, y)$ of a distribution function F and an outcome y. It is assumed that this rule is used to judge the quality or success of forecasts in the form of distribution functions. If $F_1, \ldots, F_N$ is a series of realizations of predictive distribution functions, and $y_1, \ldots, y_N$ is a series of actual outcomes, then the forecast is more successful the greater is its average score given by

$$\frac{1}{N}\sum_{i=1}^{N} S(F_i,\, y_i).$$

For a scoring rule $S(F,\, y)$ let $S(F_2, F_1)$ be the expected score of a forecast distribution F_2 under the assumption that the outcome Y is distributed as F_1. That is,

$$S(F_2,\, F_1) = \mathsf{E}S(F_2,\, Y) \qquad \text{for } Y \sim F_1.$$

By definition the scoring rule S is *proper*, if

$$S(F_1,\, F_1) \geq S(F_2,\, F_1),$$

and it is *strictly proper*, if the inequality is strict for $F_2 \neq F_1$. If the forecast is assessed according to a proper scoring rule, then the forecaster cannot expect to benefit by cheating and reporting forecast distribution which he believes to be incorrect. A detailed review of this topic can be found in Gneiting and Raftery (2007) and Bröcker and Smith (2007).

When forecasts are in the form of distribution functions it is logical to base forecast evaluation on the notion of a proper scoring rule, because it is closely related to the maximization of expected utility by the forecast user. Indeed, define a scoring rule S as the utility of an outcome y under the best action $a(F)$:

$$S(F,\, y) = u(y,\, a(F)).$$

Such a utility-based scoring rule is proper since

$$S(F_1,\, F_1) = \mathsf{E}u(Y,\, a(F_1)) \geq \mathsf{E}u(Y,\, a(F_2)) = S(F_2,\, F_1) \qquad \text{for } Y \sim F_1$$

(see, for example, Gneiting and Raftery, 2007). Therefore, when analyzing the quality of probabilistic forecasts one can focus on proper scoring rules and abstract from the implicit expected utility maximization.

A proper scoring rule was first proposed for discrete outcomes by G. Brier (Brier, 1950). Let a variable Y take on k values $(1, \ldots, k)$, $\pi_j = \mathsf{P}_F\{Y = j\}$ be the probabilities that the predictive distribution F associates with the events $Y = j$ and $\mathrm{I}\{A\}$ be the indicator of an event (condition) A. Then the *Brier scoring rule* (also called the quadratic scoring rule) is given by

$$S(F,\, y) = -\sum_{j=1}^{k} \left(\pi_j - \mathrm{I}\{y = j\}\right)^2.$$

In econometrics the most widely used is the *logarithmic scoring rule* for density forecasts. If the predictive distribution F is absolutely continuous, then $F'(y)$ is the corresponding density function and the logarithmic scoring rule is defined by

$$S(F, y) = \log F'(y).$$

For a discrete distribution

$$S(F,y) = \log \pi_y.$$

There is an obvious close relation between the logarithmic scoring rule and the log-likelihood function.

A less popular, but quite natural scoring rule is the *continuous ranked probability score* (CRPS). Its original definition is

$$S(F,\, y) = -\textstyle\int_{-\infty}^{\infty} \left(F(t) - \mathrm{I}\{t \le y\}\right)^2 dt.$$

It depends on an integral of squared distance in the L^2 metric between forecast distribution function and the empirical distribution function for a single observation y. Gneiting and Raftery (2007) propose an alternative form of this rule:

$$S(F,y) = \frac{1}{2}\mathsf{E}|Y - Y'| - \mathsf{E}|Y - y| \qquad \text{for independent } Y \sim F \text{ and } Y' \sim F.$$

This scoring rule is appealing, because it can be viewed as a generalization of the absolute distance loss which is a popular criterion for evaluation of point forecast.

2.2 Calibration

The idea of calibration It is important for a probabilistic forecast to be calibrated (Diebold, Hahn and Tay, 1999; Gneiting et al., 2007). Calibration means good conformity between the probabilistic forecasts and the actual behavior of the predicted variable. In practice probabilistic forecasts are not always well-calibrated. In particular, a common phenomenon is overconfidence (Lichtenstein et al., 1982). For example, consider a situation where people report central 90% forecast intervals. The intention is that for such an interval 5% of outcomes should be below the lower bound of the interval, 5% above the upper bound and 90% within the interval. In practice it may well be that only 50% of the outcomes fall in the interval. One can also observe downward or upward biases in forecast distributions. For example, an upward bias (in the case of a central 90% interval) may result in a situation when 9% of the outcomes are above the upper bound and only 1% of outcomes are below the lower bound.

The cause of such inconsistency can be that forecast success is judged not by proper scoring rules, but by some other criteria. For example, to appear a more skillful forecaster a person can report a forecast interval that is too narrow, thus exaggerating his expertise. It can also happen that a forecaster has an incentive to report good news (for example, low forecasts of unemployment and high forecasts of GDP growth) rather than bad ones. Some people, on the contrary, tend to overestimate probabilities of adverse events; perhaps they consciously or unconsciously aspire to be in a situation when they could say “Well, I told you that there will be a crisis.”

Even if subjective judgments do not play a great role, as in forecasting using econometric models, miscalibrated forecasts are still very common, as all models are imperfect to varying degrees. To improve methods and models used for making forecasts we need to be able to diagnose miscalibration. This can help to correct forecasts and make them “more calibrated”.

PIT-calibration To diagnose forecast calibration in the case of density forecasting one can use the *probability integral transform* values (PIT values).[2] This indicator is the

[2] The notion of PIT can be extended to arbitrary distributions by introducing randomization (Brockwell, 2007).

one that is used most often for calibration diagnostics in econometrics (e.g. Diebold et al., 1998; Mitchell and Wallis, 2011). If a forecast distribution F is reported while the actual outcome is y then the corresponding PIT value is defined as $F(y)$. The PIT values $P = F(Y)$ of a well-calibrated forecast of Y should be uniformly distributed on the interval $[0, 1]$, i.e. $P \sim U[0, 1]$. For example, if $F(M) = \frac{1}{2}$, then M is the median of the forecast distribution and under adequate calibration probability that Y is less than M is 50%. Consequently the probability that P does not exceed $F(M) = \frac{1}{2}$ should also be equal to 50%. The same is true for other quantiles, i.e. the probability of the event $P \leq \alpha$ must be equal to α. In this paper we will call such calibration *PIT-calibration*.[3] PIT-calibration can be assessed, for example, with the help of a histogram of the PIT values on the $[0, 1]$ interval. The histogram should be almost flat (Diebold et al., 1998; Gneiting et al., 2007; Mitchell and Wallis, 2011).

Marginal calibration It can be seen that the concept of PIT-calibration relates to interval forecasting and quantile forecasting (e.g. value-at-risk forecasting). This concept assumes that probabilities are fixed while the bounds are reported by the forecaster. A reversed situation is when bounds are fixed while the forecaster reports probabilities as in the Survey of Professional Forecasters. A calibrated forecast must supply probabilities which are in accordance with the true ones (cf. Clements, 2004). Formally, a forecast F is *marginally calibrated* if $\mathsf{E}F(y) = G(y)$ for any real y, where $G(y)$ is the unconditional distribution function of Y. The term was proposed in Gneiting et al. (2007).

Marginal calibration implies that point forecasts derived from a probabilistic forecast must be unbiased if the point forecast corresponds to a distribution moment such as the mean. For the mean we must have $\mathsf{E}[\mathrm{mean}(F)] = \mathsf{E}Y$.

Auto-calibration In density forecasting situation PIT-calibration and marginal calibration are different concepts. Neither of them generalizes the other one.[4] A concept which subsumes both PIT-calibration and marginal calibration can be called *auto-calibration*. A forecast F is auto-calibrated if $F(y) = G_F(y)$, where $G_F(y) = G(y|F)$ is the distribution function of Y conditional on F. Here the information used to assess forecast calibration is the forecast itself. The easiest way to understand this property is to consider forecasting of a dichotomous 0/1 variable. Among the cases in which the forecast assigns probability π to the event $Y = 1$ the event must occur with this same probability: $\mathsf{P}(Y = 1|\pi) = \pi$.[5]

Calibration with respect to an information set Previous arguments do not pay enough attention to the conditional nature of calibration and to efficient use of information available to forecasters.[6]

[3] In Gneiting et al. (2007) this aspect of calibration is called probabilistic calibration.

[4] Counterexamples can be found in Gneiting et al. (2007) and in the autoregression example below taken from Mitchell and Wallis (2011) (Combo and Unfocus forecasts).

[5] Galbraith and van Norden (2011), p. 1042: “For example, if we forecast that the probability of a recession beginning in the next quarter is 20%, and of all occasions on which we make this forecast the proportion in which a recession actually begins is 20%, and if this match holds for all other possible predicted probabilities, then the forecasts are correctly calibrated”.

[6] Clements and Taylor (2003), p. 446: “Evaluating probability forecasts by calibration ignores the conditional aspect”.

By using for a forecast F based on some available information $\mathcal{A}$ analysis which is conditional on $\mathcal{A}$ we can strengthen the definition of calibration.[7] A forecast F based on the information set $\mathcal{A}$ is *calibrated with respect to* $\mathcal{A}$ if $F(y) = G_{\mathcal{A}}(y)$, where $G_{\mathcal{A}}(y) = G(y|\mathcal{A})$ is the distribution function of Y conditional on $\mathcal{A}$.[8] Such a forecast can be called *ideal* (of all forecasts based on $\mathcal{A}$). A forecast which is calibrated with respect to an information set is always auto-calibrated and hence both PIT-calibrated and marginally calibrated.[9]

In the case of density forecasting the PIT value $P = F(Y|\mathcal{A})$ of a probabilistic forecast F which is ideal with respect to $\mathcal{A}$ is not only uniformly distributed on the interval $[0, 1]$, but also uniformly distributed conditionally on $\mathcal{A}$: $P|\mathcal{A} \sim U[0, 1]$. Similarly, point forecasts derived from a probabilistic forecast must be not only unconditionally unbiased, but also unbiased conditionally on $\mathcal{A}$. These properties of ideal forecasts are discussed further in Subsection 3.2.

An important property of an ideally calibrated forecast is that it achieves the maximum expected score when the scoring rule used is proper:[10]

$$\mathsf{E}S(G_{\mathcal{A}}, Y) \geq \mathsf{E}S(F, Y).$$

This is the reason for calling a well-calibrated forecast ideal (or efficient). Moreover, under appropriate additional conditions the inequality here is strict if the scoring rule S is strictly proper. Thus, if a forecast is miscalibrated, then there is a potential for its improvement as measured by the mean score according to a proper scoring rule. In a certain sense the concept of calibration is intrinsically based on proper scoring rules and score maximization.

Independence and uniformity of the PIT values If a sequence of one-step density forecast of a time series Y_t is made from the full history of the same series, then calibration is usually tested by analyzing the resulting series of PIT values. Consider a sequence F_t of probabilistic forecasts of a univariate time series Y_t based on its own previous history $\sigma(Y_1, \ldots, Y_{t-1})$, $T = 1, 2, \ldots$ (For $t = 1$ the forecast is unconditional). Define the corresponding PIT values as

$$P_t = F_t(Y_t), \qquad t = 1, 2, \ldots.$$

The sequence of forecasts is correctly calibrated if and only if the PIT values P_t are independent and distributed as $U[0, 1]$ (cf. Diebold et al., 1998).

[7] Formally, the random element F is $\mathcal{A}$-measurable.

[8] The definition was proposed independently in Gneiting and Ranjan (2011). It is also similar to the definition of interval forecast efficiency with respect to the information set in Christoffersen (1998).

[9] In general, when F is based on $\mathcal{A}$, we have $\sigma(F) \subset \mathcal{A}$ and $\mathsf{E}[G_{\mathcal{A}}(y)|F] = G_F(y)$. Further, if $F = G_{\mathcal{A}}$ (F is ideal) we have $\mathsf{E}[G_{\mathcal{A}}(y)|F] = \mathsf{E}[G_{\mathcal{A}}(y)|G_{\mathcal{A}}] = G_{\mathcal{A}}(y)$ and thus $F(y) = G_{\mathcal{A}}(y) = G_F(y)$ (auto-calibration).

Gneiting and Ranjan (2011) note that the ideal forecast is both marginally calibrated and probabilistically calibrated (PIT-calibrated). An outline of the proofs is as follows. Since $\mathsf{E}[G_{\mathcal{A}}(y)] = G(y)$ for any y, under ideal calibration with respect to $\mathcal{A}$ we have $\mathsf{E}F(y) = G(y)$ (marginal calibration). Further, as explained below, under ideal calibration with respect to $\mathcal{A}$, we have $F(Y)|\mathcal{A} \sim U[0, 1]$ which implies unconditional uniformity $F(Y) \sim U[0, 1]$ (PIT-calibration).

Since auto-calibration of F means that F is ideally calibrated with respect to $\sigma(F)$, it follows that auto-calibration implies both marginal calibration and PIT-calibration, as was stated above.

[10] Diebold et al. (1998), p. 866: "…If a forecast coincides with the true data generating process, then it will be preferred by all forecast users, regardless of loss function." See also Granger and Pesaran (2000).

Independence of the PIT values can be judged, for example, by the autocorrelation function of the PIT values and their transformations (Diebold et al., 1998). Independence implies the absence of serial correlation.

In general, uniformity and independence of the PIT values do not indicate ideal forecast, since the forecast can incorporate extraneous noise. In addition, this property does not hold for multi-step forecasts and forecasts using revised real-time data. Furthermore, independence and uniformity of the PIT values is necessary, but not sufficient for the ideal calibration of forecasts of a series which can use the history of some other series.

Refining the notion of calibration Ideal calibration refers to some given information set $\mathcal{A}$. However, in a forecast evaluation situation one should distinguish (at least) two different parties: the forecaster and the individual who evaluates the forecast. The second party will be called the examiner here. The information sets of the forecaster and the examiner can be distinct, say, $\mathcal{A}_f$ and $\mathcal{A}_e$. Then the notion of ideal calibration is ambiguous without specifying the information set. The forecast which is calibrated with respect to $\mathcal{A}_f$ can be not calibrated with respect to $\mathcal{A}_e$ and vice versa.

If the forecaster possesses some information which is not available to the examiner then the examiner can potentially derive some new information from the forecast. Thus, the relevant information set in this case combines $\mathcal{A}_e$ with the information delivered by the forecast. This suggests an extension of the notions of auto-calibration and ideal calibration: a forecast F is calibrated from the examiner's point of view if it coincides with $G(\cdot|\mathcal{A}_e, F)$, which is the conditional distribution of Y given $\mathcal{A}_e$ and F.

The information which the forecaster reports to the examiner can be limited. If the forecaster reports only quantiles and forecast intervals then we deal with an extended version of PIT-calibration. If the forecaster reports only probabilities for some fixed intervals then we deal with an extended version of marginal calibration.

This considerations do not undermine the general framework, but provide a refinement. For example, to test forecast calibration the examiner can still use the relevant moment conditions discussed in Section 3.

2.3 Forecast calibration and forecast sharpness

Forecast sharpness is a characteristic which reflects the degree of forecast definiteness, the concentration of a forecast distribution (Gneiting et al., 2007). One way to visualize this feature is to consider the forecast interval corresponding to a forecast distribution. For example, if an expert declares that by the end of the year EUR/USD rate will surely be in the range 1.39–1.41, it would be a rather sharp prediction. If he declares an interval 0.5–2.0, then the forecast is very vague. On the one hand, the user of the forecast would prefer to have a very sharp prediction, but on the other hand correct calibration is also important. When producing too sharp a forecast the expert risks to exaggerate the extent of his confidence which can lead to surprises for the forecast user (often not very pleasant).

In Gneiting et al. (2007) a conjecture was stated that the problem of finding a good forecast can be viewed as the problem of maximizing sharpness subject to calibration. It can be shown that the conjecture is actually true provided that a vague "calibration" notion is replaced by auto-calibration.

First, for a proper scoring rule $S(F, F)$ can be viewed as a measure of sharpness of a forecast F. For a proper scoring rule $-S(F, F)$ is a concave[11] function of F and thus,

[11] Function $S(F_1, F_2)$ is linear in the second argument as the Stieltjes integral with respect to this

according to DeGroot (1962), can be viewed as a measure of uncertainty of a probability distribution F. For the logarithmic scoring rule $-S(F,F)$ is the familiar Shannon's entropy measure.

Second, for an auto-calibrated forecast we have $\mathsf{E}S(F,Y) = \mathsf{E}S(F,F)$, i.e. the expected score of an auto-calibrated forecast equals its expected sharpness.[12]

This means that auto-calibrated forecasts can be compared on the basis of the levels of their expected sharpness. The ideal forecast is the sharpest of all auto-calibrated forecasts, because it is characterized by the greatest expected score.

Another intuitively expected property of well-calibrated forecasts is that the more complete information has the forecaster, the sharper is the forecast which he can potentially produce. Suppose $\mathcal{A}_1$ is a "richer" information set than $\mathcal{A}_2$, that is, $\mathcal{A}_1$ contains all the information of $\mathcal{A}_2$ and maybe some additional useful information (formally, $\mathcal{A}_2 \subset \mathcal{A}_1$). Let $G_1 = G_{\mathcal{A}_1}$ be the ideal forecast based on $\mathcal{A}_1$ and $G_2 = G_{\mathcal{A}_2}$ the ideal forecast based on $\mathcal{A}_2$. Then[13]

$$\mathsf{E}S(G_1,Y) = \mathsf{E}S(G_1,G_1) \geq \mathsf{E}S(G_2,Y) = \mathsf{E}S(G_2,G_2).$$

If a forecast is not auto-calibrated, then its sharpness can be deceiving. Let d denote a divergence indicator (generalized distance) between distributions F_1 and F_2 defined as

$$d(F_2,F_1) = S(F_1,F_1) - S(F_2,F_1).$$

The divergence $d(F_2,F_1)$ is non-negative, if the rule S is proper. It is zero when the two distributions coincide. For the logarithmic scoring rule d is the Kullback–Leibler distance. In general the expected score of a (possibly miscalibrated) forecast F can be decomposed as follows:

$$\mathsf{E}S(F,Y) = \mathsf{E}S(G_F,G_F) - \mathsf{E}d(F,G_F),$$

where G_F is the conditional distribution function of Y given F. The first term can be interpreted as the expected sharpness of the forecast G_F, which is a "recalibrated" version of forecast F, while the second term relates to the divergence between F and G_F, i.e. it is a measure of miscalibration of the forecast F with respect to the information contained in itself.[14]

The principle of maximizing sharpness subject to calibration which was considered here is difficult to apply in practice, because achieving perfect auto-calibration of a forecast may prove too challenging. However, this principle provides a useful insight into the essence of probabilistic forecasting. In particular, it is clear that the advantage of using proper scoring rules is that they provide the right balance of sharpness and calibration in forecast comparison. If other—not proper—scoring rules are used for forecast evaluation, then the forecaster would have an incentive to report miscalibrated, for example, too sharp forecasts.

argument. Therefore $S(F_\alpha,F_\alpha) = \alpha S(F_\alpha,F_1) + (1-\alpha)S(F_\alpha,F_2) \leq \alpha S(F_1,F_1) + (1-\alpha)S(F_2,F_2)$ for $F_\alpha = \alpha F_1 + (1-\alpha)F_2$ and $\alpha \in [0,1]$.

[12] Indeed, for a non-random distribution function H define $T(H) = \mathsf{E}[S(H,Y)|F] = S(H,G_F)$. Then $T(H) = S(H,F)$ since $F = G_F$ (we assume that F is auto-calibrated). By the substitution property of the conditional expectation $T(F) = S(F,F)$ since F is $\sigma(F)$-measurable. Taking unconditional expectation yields $\mathsf{E}[S(F,Y)] = \mathsf{E}[T(F)] = \mathsf{E}[S(F,F)]$.

[13] See Holzmann and Eulert (2011) for a proof. Similar results for discrete outcome can be found in DeGroot and Fienberg (1983) and Bröcker (2009).

[14] This partitioning for the dichotomous outcomes and the Brier score was developed in Sanders (1963). Bröcker (2009) extended it to the case of an arbitrary discrete distribution and an arbitrary proper scoring rule.

3 Moment-based calibration testing

3.1 The general idea of moment-based calibration testing

In practice it is convenient to express the calibration conditions in the form of moment conditions. A realization of a complete probabilistic forecast defines a probability measure for outcomes, which can be used to calculate various moments. If the forecast is calibrated and coincides with a conditional distribution function, then the corresponding conditional moments can be expressed in terms of the moments calculated from the generated probability measures. One can replace theoretical moments by sample ones based on a series of forecasts and see how far the result is from what should be in theory. This allows to develop various types of diagnostic tests for forecast calibration. Many of the tests and criteria of calibration/efficiency developed in the literature can be shown to fall within this approach.

Suppose that in theory under calibration the expectation of z must be zero: $\mathsf{E}z = 0$. We can obtain the values of z for a series of realizations of forecast functions $F_1, \ldots, F_N$ and a series of outcomes $y_1, \ldots, y_N$ and calculate the corresponding sample moment $\bar{z} = \sum_{i=1}^{N} z_i/N$. If $\bar{z}$ is far from zero, then we can conclude that the forecast is miscalibrated.

To test the moment conditions we can use the usual t-ratios $\bar{z}/se(\bar{z})$. The most subtle aspect here is adequate calculation of the standard error $se(\bar{z})$. In the examples below the usual heteroskedasticity and autocorrelation consistent (HAC) standard errors are used. If this is done correctly and the forecast is well-calibrated, then this statistic is asymptotically distributed as $N(0, 1)$. An extension to the multivariate case—simultaneous testing of several moment conditions—is straightforward and is familiar from the GMM framework: a t-ratio is replaced by a quadratic form and the distribution is chi-square. For the orthogonality conditions discussed below testing could be conveniently done by means of ordinary F-statistics from auxiliary regressions provided that heteroskedasticity and serial correlation are not an issue. However, auxiliary regressions are by no means necessary for testing of moment conditions.[15]

3.2 Testing for PIT-calibration

PIT-calibration of a density forecast can be tested by comparing sample moments of the PIT values with the corresponding moments of the $U[0, 1]$ distribution. Consider a function $k = k(p)$ taking a probability $p \in [0, 1]$ as its argument and define

$$\kappa = \mathsf{E}k(P) \quad \text{for } P \sim U[0, 1].$$

If a forecast F is PIT-calibrated, then

$$\mathsf{E}k(F(Y)) - \kappa = 0.$$

In particular, using this notation we can write the condition that the probability of the event $F(Y) \leq \alpha$ is α as mentioned above in the discussion of PIT-calibration. To do this, take

$$k = \mathrm{I}\{p \leq \alpha\}, \qquad \kappa = \alpha. \tag{1}$$

[15] For example Berkowitz (2001) and Clements and Taylor (2003) propose likelihood ratio tests based on auxiliary regressions. The tests are not robust to serial correlation. In general this would not lead to the asymptotic chi-square distribution.

Next, consider a central forecast interval with the coverage probability β. If F is a forecast distribution function, then the interval is of the form

$$[F^{-1}(1/2-\beta/2), F^{-1}(1/2+\beta/2)].$$

Under PIT-calibration the probability that Y belongs to this interval is equal to β. An outcome y is in this interval if and only if the PIT value $F(y)$ is in the interval $C_\beta = [0.5-0.5\beta, 0.5+0.5\beta]$. Therefore, here one can take

$$k = \mathrm{I}\{p \in C_\beta\}, \qquad \kappa = \beta. \tag{2}$$

If $F(Y)$ is distributed as $U[0,1]$, then the inverse normal transform (INT) of this variable has the standard normal distribution:

$$\Phi^{-1}(F(y)) \sim N(0,1),$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. Instead of moments of the PIT values one can use moments of the INT values. For example, the INT values must have zero mean:

$$k = \Phi^{-1}(F(Y)), \qquad \kappa = 0.$$

3.3 Testing for marginal calibration

Let $m(y)$ be some function of an outcome y and denote

$$\mu(F) = \mathsf{E}m(Y) \quad \text{for } Y \sim F.$$

Then under marginal calibration

$$\mathsf{E}m(Y) - \mu(F) = 0,$$

that is, $\mu(F)$ is an unbiased forecast of $m(Y)$. In particular, $m = y$, $\mu = \mathrm{mean}(F)$ allows to test for mean unbiasedness of F.

For example, if F is the distribution function of the outcome Y, then $F(y_0)$ is the probability of the event $Y \leq y_0$. The corresponding moment condition can be tested with

$$m = \mathrm{I}\{y \leq y_0\}, \quad \mu = F(y_0),$$

which directly corresponds to the definition of marginal calibration. More generally, if C is some fixed interval for the outcome, then we can use

$$m = \mathrm{I}\{y \in C\}$$

and

$$\mu = \mathsf{P}(Y \in C) \quad \text{for } Y \sim F.$$

3.4 Testing for auto-calibration

From the theory of point forecasting it is known that the expectation conditional on the information set $\mathcal{A}$ is the forecast which is optimal in mean-square sense among the forecast based on $\mathcal{A}$ (e.g. Bierens, 2004, pp. 80–81). This forecast satisfies an orthogonality condition: the prediction error is uncorrelated with any random variable based on $\mathcal{A}$.[16] There are also extensions to the case of general cost functions (e.g. Granger, 1999). In Mitchell and Wallis (2011) an idea was put forward that calibration of probabilistic forecasts can be tested by verifying similar orthogonality conditions. We demonstrate that this idea lends itself to further generalization.

Let k and κ have the same meaning as above and let $l(F)$ be a function of a distribution function F. If forecast F is auto-calibrated, then we have a conditional restriction on the moments

$$\mathsf{E}[k(F(Y)) - \kappa | F] = 0,$$

which implies the orthogonality condition

$$\mathsf{E}[(k(F(Y)) - \kappa)\, l(F)] = 0 \tag{3}$$

for any function l. This means that any function k of the PIT values is not correlated with any function l of the forecast distribution function. Here l can be some characteristics of the forecast distribution such as the mean or median.

By the same logic, for m and μ having the same meaning as above

$$\mathsf{E}[m(Y)|F] = \mu(F)$$

and

$$\mathsf{E}[(m(Y) - \mu(F))\, l(F)] = 0.$$

This means that the forecast error for the point forecast $\mu(F)$ derived from F is not correlated with any other feature l of F. An example of a test using this type of orthogonality conditions can be found in Clements (2004), where it was applied to evaluation of the SPF probabilistic forecasts.

To test auto-calibration of a forecast one can also use more general moment conditions. Consider a function $r = r(y, F)$ taking an outcome y and a distribution function F as its arguments and denote

$$\rho(F) = \mathsf{E}r(Y, F) \quad \text{for } Y \sim F.$$

Auto-calibration of a probabilistic forecast F is equivalent to the condition

$$\mathsf{E}r(Y, F) = \mathsf{E}\rho(F)$$

for any r.

For example, consider some interval $C = C(F)$ defined for a distribution function F and its theoretical coverage

$$a = a(F) = \mathsf{P}(Y \in C) \quad \text{for } Y \sim F.$$

Then we can test auto-calibration with $r = \mathrm{I}\{y \in C\}$ and $\rho = a$, where both C and a can vary with F (while under PIT-calibration α is fixed and under marginal calibration C is fixed).

[16] These conditions were utilized in the rational expectations literature. Shiller (1978), p. 7: "…Expected forecast errors conditional on any subset of the information available when the forecast was made, are zero… Hence, the forecast error … is uncorrelated with any element of I_t [the set of public information available at time t]".

3.5 Testing for calibration with respect to an information set

Ideal calibration of F with respect to $\mathcal{A}$ is a stronger property and requires

$$\mathsf{E}[r(Y,F)|\mathcal{A}] = \rho(F)$$

for any r. This conditional property is equivalent to unconditional orthogonality

$$\mathsf{E}[(r(Y,F) - \rho(F))W] = 0$$

for any W depending on $\mathcal{A}$.[17]

In particular, for the PIT values we have

$$\mathsf{E}[k(F(Y))|\mathcal{A}] = \kappa$$

and the corresponding orthogonality condition is given by

$$\mathsf{E}[(k(F(Y)) - \kappa)W] = 0.$$

For example, one can define

$$k = \mathrm{I}\{p \leq \alpha\}, \qquad \kappa = \alpha$$

and test ideal calibration with respect to $\mathcal{A}$ by testing the lack of correlation between the indicator variable $\mathrm{I}\{F(Y) \leq \alpha\}$ for $\alpha \in (0,1)$ and any function W depending only on the available information $\mathcal{A}$.[18] Note that if

$$\mathsf{E}[\mathrm{I}\{F(Y) \leq \alpha\}|\mathcal{A}] = \alpha$$

for any α then F is calibrated with respect to $\mathcal{A}$.[19]

Marginal calibration can also be strengthened along this lines:

$$\mathsf{E}[m(Y)|\mathcal{A}] = \mu(F)$$

and

$$\mathsf{E}[(m(Y) - \mu(F))W] = 0.$$

Thus, μ must be the conditional expectation of $m(Y)$ given $\mathcal{A}$. Consequently, μ must be an unbiased point forecast of m and the prediction error must be uncorrelated with any variable W constructed from the available information $\mathcal{A}$. For example, one can take $m = \mathrm{I}(y \leq y_0)$, $\mu = F(y_0)$ for some fixed y_0. If $\mathsf{P}(Y \leq y_0|\mathcal{A}) = \mathsf{E}[\mathrm{I}(Y \leq y_0)|\mathcal{A}] = F(y_0)$ for any real y_0 then $F = G_{\mathcal{A}}$ by the definition of $G_{\mathcal{A}}$, i.e. F is calibrated with respect to $\mathcal{A}$.

[17] That the two conditions are equivalent is a well-known property of the conditional expectation. Let $E = r(Y,F) - \rho(F)$. From the conditional unbiasedness $\mathsf{E}[E|\mathcal{A}] = 0$ when W is $\mathcal{A}$-measurable we have $\mathsf{E}[EW|\mathcal{A}] = \mathsf{E}[E|\mathcal{A}]W = 0$, and thus $\mathsf{E}[EW] = 0$. To see that the conditional unbiasedness follows from the unconditional orthogonality let $W = \mathrm{I}(A)$ for an arbitrary $A \in \mathcal{A}$.

[18] Christoffersen (1998) proposed a similar condition for testing for conditional coverage of an interval forecast.

[19] For an invertible non-random distribution function H we have $\mathsf{P}(H(Y) \leq \alpha|\mathcal{A}) = \mathsf{E}[\mathrm{I}\{H(Y) \leq \alpha\}|\mathcal{A}] = \mathsf{E}[\mathrm{I}\{Y \leq H^{-1}(\alpha)\}|\mathcal{A}] = G_{\mathcal{A}}(H^{-1}(\alpha))$. Thus, by the substitution property of the conditional expectation when F is $\mathcal{A}$-measurable $\mathsf{P}(F(Y) \leq \alpha|\mathcal{A}) = \mathsf{E}[\mathrm{I}\{F(Y) \leq \alpha\}|\mathcal{A}] = G_{\mathcal{A}}(F^{-1}(\alpha))$. Further, $G_{\mathcal{A}}(F^{-1}(\alpha)) = \alpha$ for each α is equivalent to $F = G_{\mathcal{A}}$.

To generalize this approach consider a function $g(y, F, w)$ taking an outcome y, a distribution function F and some additional variable w as its arguments. Define

$$\gamma(F, w) = \mathsf{E}[g(Y, F, w)] \quad \text{for } Y \sim F.$$

Calibration of a forecast F with respect to $\mathcal{A}$ is equivalent to the moment condition

$$\mathsf{E}[g(Y, F, W)|\mathcal{A}] = \gamma(F, W)$$

or

$$\mathsf{E}g(Y, F, W) = \mathsf{E}\gamma(F, W)$$

for any such g and any W depending on $\mathcal{A}$.[20]

Consider an example of conditions of this more general type. The idea is to test calibration of one forecasting method against another one.[21] Suppose that we want to test whether F_1 is well-calibrated and F_2 is an alternative forecast, both based on $\mathcal{A}$. For a proper scoring rule S and two (non-random) distribution functions F_1, F_2 define

$$g = S(F_2, Y) - S(F_1, Y).$$

Then we have

$$\gamma = S(F_2, F_1) - S(F_1, F_1).$$

Consequently, for two forecasts F_1, F_2 based on $\mathcal{A}$ under the assumption that F_1 is calibrated with respect to $\mathcal{A}$ we have

$$\mathsf{E}[S(F_2, Y) - S(F_1, Y)] = \mathsf{E}[S(F_2, F_1) - S(F_1, F_1)]. \tag{4}$$

Below we call this moment condition *relative forecast calibration* (RFC). A test which is based on it would have power against an alternative that F_2 is calibrated with respect to $\mathcal{A}$, because then

$$\mathsf{E}[S(F_2, F_2) - S(F_1, F_2)] + \mathsf{E}[S(F_1, F_1) - S(F_2, F_1)] > 0,$$

if the scoring rule is strictly proper.

A pair of reciprocal RFC-based tests (F_1 against F_2 and F_2 against F_1) can help to judge possible gains from combining two forecasts.

[20] Denote $\gamma_0(\hat{H}, H, w) = \mathsf{E}[g(Y, H, w)]$ for $Y \sim \hat{H}$, where $\hat{H}, H$ are non-random distribution functions. First, by the properties of the conditional expectation, conditional distribution function $G_{\mathcal{A}}$ can be used to calculate the conditional expectation with respect to $\mathcal{A}$: $\mathsf{E}[g(Y, H, w)|\mathcal{A}] = \gamma_0(G_{\mathcal{A}}, H, w)$. Second, by the substitution property, the $\mathcal{A}$-measurable random elements F and W can be treated as fixed inside the conditional expectation with respect to $\mathcal{A}$ and thus $\mathsf{E}[g(Y, F, W)|\mathcal{A}] = \gamma_0(G_{\mathcal{A}}, F, W)$. Third, if $F = G_{\mathcal{A}}$, then $\gamma_0(G_{\mathcal{A}}, F, W) = \gamma_0(F, F, W) = \gamma(F, W)$. Consequently, when F is calibrated with respect to $\mathcal{A}$, we have $\mathsf{E}[g(Y, F, W)|\mathcal{A}] = \gamma(F, W)$. The converse can be obtained by setting $W = \mathrm{I}(A)$, $g = \mathrm{I}(y \leq y_0)w$ and $\gamma = F(y_0)w$ for an arbitrary $A \in \mathcal{A}$ and an arbitrary real y_0.

[21] This example parallels the equal forecast accuracy test based on the difference of the logarithmic scores (or Kullback–Leibler information criterion, KLIC; see Amisano and Giacomini, 2007; Mitchell and Wallis, 2011).

Table 1: Definitions of six forecasts of AR(2)

Ideal	$N(\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2},\, 1)$
Climt	$N(0,\, \sigma_Y^2)$
AR1	$N(\rho_1 Y_{t-1},\, \sigma_1^2)$
AR2	$N(\rho_2 Y_{t-2},\, \sigma_2^2)$
Combo	$0.5\, N(\rho_1 Y_{t-1},\, \sigma_1^2) + 0.5\, N(\rho_2 Y_{t-2},\, \sigma_2^2)$
Unfocus	$0.5\, N(\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2},\, 1) + 0.5\, N(\varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \tau_t,\, 1)$, where $\tau_t = -1$ or 1 with equal probabilities

Note: $\rho_1 = \varphi_1/(1-\varphi_2)$, $\rho_2 = \varphi_1\rho_1 + \varphi_2$,
$\sigma_Y^2 = 1/(1 - \varphi_1\rho_1 - \varphi_2\rho_2)$, $\sigma_1^2 = (1-\rho_1^2)\sigma_Y^2$, $\sigma_2^2 = (1-\rho_2^2)\sigma_Y^2$

4 Example: Forecasting of an autoregressive process

The first illustration of the ideas is an artificial simulation example taken from Mitchell and Wallis (2011). It relates to forecasting of AR(2) series $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$ with independent Gaussian disturbances $\varepsilon_t \sim N(0,1)$. Six forecasts are compared.

1. Ideal, the ideal forecast which takes into account all available information, i.e. Y_{t-1} and Y_{t-2};

2. Climt, the "climatological" forecast which does not use Y_{t-1} and Y_{t-2} and is represented by the unconditional distribution;

3. AR1, using only Y_{t-1};

4. AR2, using only Y_{t-2};

5. Combo, a combined forecast, which is a convex combination of AR1 and AR2 with equal weights;

6. Unfocus, an "unfocused" forecast containing extraneous noise τ_t.

Table 1 gives the corresponding formal description. The actual data-generating process is represented by the ideal forecast. In this example there is no need to estimate parameters as they are known. Here only the Case (2) of Mitchell and Wallis (2011) with $\phi_1 = 0.15$, $\phi_2 = 0.2$ is considered. The length of the series is $T = 150$ if not specified otherwise.

In Table 2 the row labeled "Expected score" shows approximate expected logarithmic scores obtained by Monte Carlo simulations. As might be expected, the ideal forecast has the highest expected score. The unfocused forecast shows the worst result, followed by the climatological forecast. Forecasts AR1, AR2 on average are worse than their combination Combo.

The expected score shows the asymptotic potential of a forecast which becomes visible when the number of observations tends to infinity. It cannot be available in a practical forecasting situation. When a series of forecasts is not very long, imperfect forecasts can obtain higher average scores than the ideal forecast.

The row labeled "% best" shows the percentage of experiments in which the corresponding model had the highest average logarithmic score when using $T = 150$ observations. The ideal forecast was the best with the probability of about 2/3. This moderately large value is explained by small values of the two autoregression coefficients ($\phi_1 = 0.15$,

Table 2: Statistics for six forecasts of AR(2)

	Ideal	Climt	AR1	AR2	Combo	Unfocus
Expected score	−1.418	−1.456	−1.438	−1.430	−1.425	−1.529
% best	66.5	1.2	6.8	12.2	13.2	0.1
% best, 1500	98.6	0.0	0.0	0.1	1.3	0.0
$z \times$ mean test	4.5	—	4.5	4.3	17.4	99.8
ΔL test vs. Ideal	—	45.7	32.5	25.4	16.9	90.9
RFC test vs. Ideal	—	92.6	78.3	62.4	25.8	100.0
RFC test vs. Climt	5.5	—	4.9	5.0	0.7	100.0
RFC test vs. AR1	4.7	62.7	—	40.2	3.3	100.0
RFC test vs. AR2	3.4	86.6	58.6	—	9.4	100.0
RFC test vs. Combo	4.2	86.8	48.6	31.3	—	100.0
RFC test vs. Unfoc	3.9	37.0	17.8	9.9	6.0	—

Note: The table is based on 5000 simulations. The figures for the tests are percentages of rejection at 5% asymptotic significance level using the standard normal quantiles. The test statistics are t-ratios with Newey–West HAC standard errors and lag truncation 4. "% best, 1500" corresponds to 10 times longer series ($T = 1500$). $S(F_2, F_1)$ functions for mixtures of normals needed for RFC statistics are computed by Monte Carlo with 100 simulations.

$\phi_2 = 0.2$), which makes the true data generating process rather close to AR1, AR2 and Combo alternatives in terms of the expected logarithmic score. With $T = 1500$ the ideal forecast dominates the other ones (see row labeled "% best, 1500"). One can see that the average score is a sensible criterion for model selection which behaves in a predictable way.

The unfocused forecast deserves special attention. Unconditionally it is PIT-calibrated and thus has PIT values distributed as $U[0,1]$ (cf. Table II of Mitchell and Wallis, 2011). Also the PIT series are serially independent (cf. Table III of Mitchell and Wallis, 2011). However, the forecast is not marginally calibrated and, as a consequence, is not auto-calibrated, which can be easily detected using a test of orthogonality between the INT values $z_t = \Phi^{-1}(F_t(y_t))$ and the mean of the forecast distribution. It is labeled "$z \times$ mean test" in Table 2. For an auto-calibrated forecast we have $\mathsf{E}[\mathrm{mean}(F_t)z_t] = 0$, which is an orthogonality condition of the form (3). The table shows that this condition is rejected in almost 100% of experiments in the case of the unfocused forecast.

Combo, which is a weighted combination of forecast AR1 and AR2, is marginally calibrated as a mixture of marginally calibrated forecasts. However, it is not PIT-calibrated (as is often the case for similarly constructed combined forecasts, cf. Gneiting and Ranjan, 2011) and, as a consequence, is not auto-calibrated. The results of $z \times$ mean test confirm that this forecast is not auto-calibrated (although the rejection rate is not very large).

For other forecasts there are no signs of miscalibration according to $z \times$ mean test. Indeed, Ideal, AR1 and AR2 forecasts are all auto-calibrated.

To test calibration of the forecasts against each other I employ a relative forecast calibration test based on the RFC property (4). It uses the logarithmic scoring rule. The test is designed as one-sided to increase its power, because the test statistic is expected to be positive in situations when the alternative forecast can potentially be used to improve the forecast tested for calibration. Notable are the results of AR1 vs. AR2 and AR2 vs.

AR1. The tests would frequently suggest the usefulness of combining the two models. The test against the ideal forecast can be seen to have high power. It can be compared to an equal predictive ability test based on the difference of the attained average scores $\Delta L_t = S(G_t, y_t) - S(F_t, y_t)$ in the spirit of Amisano and Giacomini (2007) (labeled "ΔL test vs. Ideal"). The later test is also implemented as one-sided to increase its power and make it comparable with the RFC test. The comparison clearly favors the RFC test as an instrument of identifying non-ideal forecasts.

5 Example: Forecasting a stock index

The second example is intended to show how the proposed framework can be utilized for the task of evaluating forecasts of real-world time series. The data are daily close levels of the Russian stock market index (RTSI) and span the period from 1995-09-01 to 2011-06-08. The RTSI series is brought to stationarity by computing its growth rates in percent $R_t = (\log RTSI_t - \log RTSI_{t-1}) \times 100$. This provides a series of 4,209 observations. The forecasted variable is the growth rate 10 periods ahead. Thus, at time t a forecast of $y_{t+10} = R_{t+1} + \cdots + R_{t+10} = (\log RTSI_{t+10} - \log RTSI_t) \times 100$ is obtained. (The ten-period horizon corresponds roughly to two weeks of physical time.) The following forecasts are considered.[22]

1. The historical forecast based on the full history Hist(t) uses all previously observed 10-period growth rates $y_{10}, \ldots, y_t$. The historical observations are resampled to obtain an ensemble of size 1000.

2. Hist(200) is also a historical forecast, but uses only a rolling span of the 200 most recent observations $y_{t-199}, \ldots, y_t$ and does not use resampling.

3. ES forecast is based on exponential smoothing for volatility $\sigma^2_{t+1} = (1-\delta)R^2_t + \delta\sigma^2_t$ with the decay factor $\delta = 0.95$ (RiskMetrics, 1996). The forecasting distribution is given by $N(0, 10\sigma^2_{t+1})$. The recursion for volatility starts from the sample variance of the first 200 observations.

4. GARCH forecast is based on the standard GARCH(1,1)-t model (GARCH with t distribution) with non-zero mean. The model is estimated recursively by the maximum likelihood method. The forecasting distribution is represented by an ensemble of 1000 simulated future trajectories.

All the forecasts are produced recursively for the forecasting period starting from the 200-th observation. They are compared by their observed averages of the CRPS. CRPS can be calculated in $\mathrm{O}(S \log S)$ operations needed for sorting if F_t is represented by a sample of size S (Gneiting and Raftery, 2007). One advantage of the CRPS over the logarithmic scoring rule is that it allows a forecast of a continuous random variable to be a discrete distribution, which we have here.

The following statistics are summarized in Table 3.

1. "CRPS" is the average CRPS.

[22] The RTSI series exhibits significant first-order serial correlation, but its impact on 10-period-ahead forecasts is small so the forecasts ignore it.

Table 3: Forecasts of RTSI index

	Hist(t)	Hist(200)	ES	GARCH
CRPS	–5.164	–5.168	–5.047	–5.038
Test α	–0.006	–0.003	–0.084***	0.005
	(0.019)	(0.019)	(0.020)	(0.019)
Test β	0.131***	0.039**	–0.026	0.005
	(0.018)	(0.017)	(0.015)	(0.016)
RFC test vs. Hist(t)	–	0.366***	0.261***	0.123***
		(0.084)	(0.054)	(0.037)
RFC test vs. Hist(200)	0.357***	–	0.324***	0.264***
	(0.077)		(0.083)	(0.077)
RFC test vs. ES	0.493***	0.566***	–	0.189***
	(0.059)	(0.096)		(0.048)
RFC test vs. GARCH	0.374***	0.525***	0.207***	–
	(0.046)	(0.093)	(0.047)	

Note: Newey–West HAC standard errors with lag truncation 10 are shown in brackets. Statistical significance at 5% (1%, 0.1%) level is shown by * (**, ***). RFC tests are CRPS-based.

2. "Test α" is an unconditional PIT-calibration statistic based on (1) with $\alpha = 0.5$. This statistic relates to the location as measured by the median of the forecasting distribution.

3. "Test β" is an unconditional PIT-calibration statistic based on (2) with $\beta = 0.5$. This statistic relates to the coverage of the central 50% interval derived from the forecasting distribution.

4. "RFC test vs. ⟨method⟩" is a one-sided RFC test, the same as in previous example, but based on the CRPS.

The selection of the calibration tests is somewhat arbitrary. The proposed framework allows to design many different tests, including other kinds of PIT-calibration tests. Such tests are formal counterparts of visual inspection of PIT histograms, a method which is frequently employed for evaluation of density forecasts.[23]

The example pertains to a typical practical situation when none of the compared forecasts can be called "ideal". All forecasts are not well-calibrated to different degrees (Table 3). For example, Hist(t) while having (unconditionally) correct location according to "Test α" notably lack sharpness which is signaled by "Test β": the actual values are too seldom found in the tails. ES rigidly assumes zero mean which is not corroborated by the observed data and accordingly "Test α" indicates a negative bias; note also the skew of the histogram in Figure 1(a).

In general "GARCH" looks almost like a well-calibrated forecast if calibration is judged on the basis of ordinary PIT-based criteria. The histogram of the PIT values at Figure 1(b) is not perfectly flat, but its unevenness is not very serious as confirmed by the two PIT-calibration tests from Table 3. Also there are no serious signs of autocorrelation

[23] The problem with such histograms is that they can be potentially deceptive without appropriate error bands.

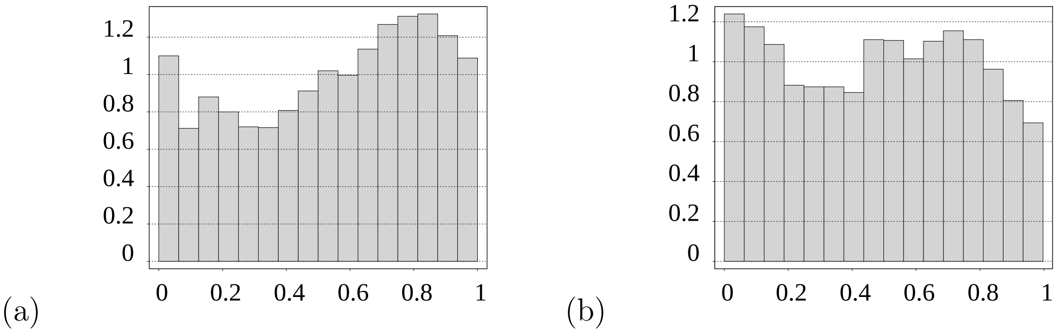

Figure 1: Histograms of the PIT values: (a) exponential smoothing, (b) GARCH

after lag 10 in both of the INT values $z_t = \Phi^{-1}(F_t(y_t))$ and the absolute INT values $|z_t|$. For example, the 11th autocorrelation coefficient is 0.078 for z_t and -0.039 for $|z_t|$, while asymptotic standard errors are 0.044 and 0.034 respectively.[24]

GARCH model is the leader in terms of the average CRPS level, followed closely by exponential smoothing. However, all of the methods in pairwise comparisons by means of RFC tests show significant miscalibration. For example, remarkably, GARCH is not able to encompass exponential smoothing, which can be considered as its "cheaper" substitute. The results show a potential for improving the forecasts by combining them.

6 Conclusion

Probabilistic forecasts provide much more information for economic decisions than conventional point forecasts, so they have good prospects. Such forecasts should be more widely adopted in practice of economic forecasting. When using probabilistic forecasts it is desirable to rely on the fundamental concepts and theoretical properties. Some of these concepts and properties are considered in this paper.

The paper argues that expected score maximization and the notion of a proper scoring rule can be viewed as the implicit basis for evaluation of probabilistic forecasts. The notion of calibration can be derived from this basis.

The paper highlights the difference between PIT-calibration and marginal calibration of density forecasts and introduces the concept of auto-calibration which generalizes them. Among other things, the concept of auto-calibration helps to derive the principle of maximizing sharpness subject to calibration from expected score maximization.

The concept of auto-calibration is further generalized by the concept of calibration with respect to an information set.

The different modes of calibration lead to different moment conditions, including orthogonality conditions. The idea of testing moment conditions leads to a general framework for calibration testing. The framework can facilitate construction of various new tests. An example is a new relative forecast calibration test (RFC test). Simulations suggest that it can have high power as a mutual calibration testing procedure for pairs of probabilistic forecasts.

[24] The Bartlett's approximation for the variance of r_{11} is used which assumes no autocorrelation after lag 10 and is given by $(1 + 2r_1^2 + \cdots + 2r_{10}^2)/T$, where r_k is the k-th autocorrelation coefficient.

References

Amisano, G. and Giacomini, R. (2007). Comparing density forecasts via weighted likelihood ratio tests, *Journal of Business and Economic Statistics* **25**(2): 177–190.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management, *Journal of Business & Economic Statistics* **19**(4): 465–474.

Bhattacharya, R. and Waymire, E. C. (2007). *A Basic Course in Probability Theory*, Springer.

Bierens, H. J. (2004). *Introduction to the Mathematical and Statistical Foundations of Econometrics*, Cambridge University Press.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability, *Monthly Weather Review* **78**(1): 1–3.

Britton, E., Fisher, P. and Whitley, J. (1998). The Inflation Report projections: Understanding the fan chart, *Bank of England Quarterly Bulletin* **38**: 30–37.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores, *Quarterly Journal of the Royal Meteorological Society* **135**(643): 1512–1519.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper, *Weather and Forecasting* **22**: 382–388.

Brockwell, A. E. (2007). Universal residuals: A multivariate transformation, *Statistics & Probability Letters* **77**: 1473–1478.

Christoffersen, P. F. (1998). Evaluating interval forecasts, *International Economic Review* **39**(4): 841–862.

Clements, M. P. (2004). Evaluating the Bank of England density forecasts of inflation, *The Economic Journal* **114**: 844–866.

Clements, M. P. and Taylor, N. (2003). Evaluating interval forecasts of high-frequency financial data, *Journal of Applied Econometrics* **18**(4): 445–456.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments, *The Annals of Mathematical Statistics* **33**(2): 404–419.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters, *Journal of the Royal Statistical Society. Series D (The Statistician)* **32**(1/2): 12–22.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management, *International Economic Review* **39**(4): 863–883.

Diebold, F. X., Hahn, J. and Tay, A. S. (1999). Multivariate density forecast evaluation and calibration in financial risk management: High-frequency returns on foreign exchange, *Review of Economics and Statistics* **81**(4): 661–673.

Diebold, F. X., Tay, A. S. and Wallis, K. F. (1999). Evaluating density forecasts of inflation: The survey of professional forecasters, *in* R. F. Engle and H. White (eds), *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, Oxford University Press, Oxford, pp. 76–90.

Engelberg, J., Manski, C. F. and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters, *Journal of Business and Economic Statistics* **27**(1): 30–41.

Galbraith, J. W. and van Norden, S. (2011). Kernel-based calibration diagnostics for recession and inflation probability forecasts, *International Journal of Forecasting* **27**(4): 1041–1057.

Gneiting, T. (2011). Making and evaluating point forecasts, *Journal of the American Statistical Association* **106**(494): 746–762.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness, *Journal of the Royal Statistical Society: Series B* **69**: 243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association* **102**: 359–378.

Gneiting, T. and Ranjan, R. (2011). Combining predictive distributions, *ArXiv preprint.* `arXiv:1106.1638v1 [math.ST]`.

Granger, C. W. J. (1999). Outline of forecast theory using generalized cost functions, *Spanish Economic Review* **1**: 161–173.

Granger, C. W. J. and Pesaran, M. H. (2000). A decision-theoretic approach to forecast evaluation, *in* W.-S. Chan, W. K. Li and H. Tong (eds), *Statistics and Finance: An Interface*, Imperial College Press.

Holzmann, H. and Eulert, M. (2011). The role of the information set for forecasting — with applications to risk management. (unpublished).

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980, *in* D. Kahneman, P. Slovic and A. Tversky (eds), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, UK, pp. 306–334.

Mitchell, J. and Wallis, K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness, *Journal of Applied Econometrics,* **26**(6): 1023–1040.

Murphy, A. H. and Winkler, R. L. (1984). Probability forecasting in meteorology, *Journal of the American Statistical Association* **79**(387): 489–500.

National Research Council (U.S.) (2006). *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*, National Academies Press.

Pesaran, M. H. and Skouras, S. (2002). Decision-based methods for forecast evaluation, *in* M. P. Clements and D. F. Hendry (eds), *A Companion to Economic Forecasting*, Blackwell Publishing, chapter 11, pp. 241–267.

RiskMetrics (1996). Riskmetrics(TM) — Technical document (4th ed.), *Technical report*, J. P. Morgan/Reuters‘.

Sanders, F. (1963). On subjective probability forecasting, *Journal of Applied Meteorology* **2**: 191–201.

Shiller, R. J. (1978). Rational expectations and the dynamic structure of macroeconomic models: A critical review, *Journal of Monetary Economics* **4**(1): 1–44.